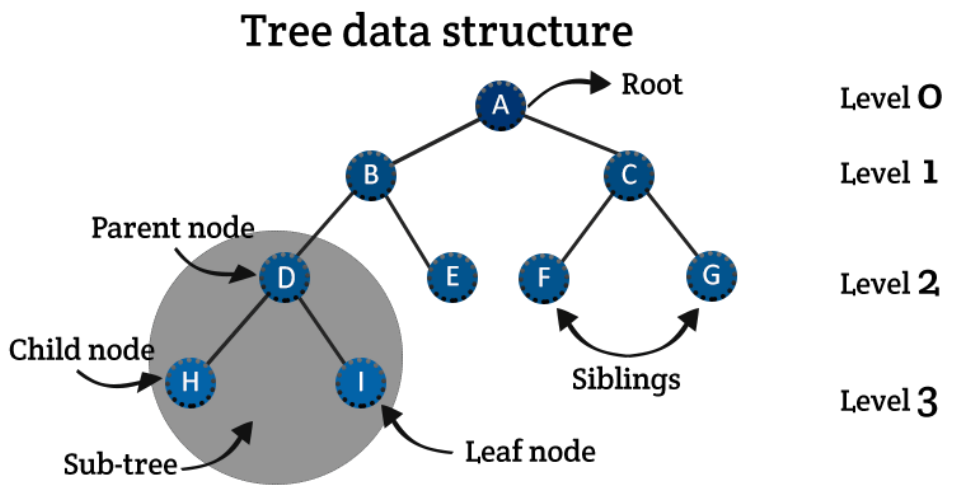
Decision Tree Algorithm

The **decision tree** is a very specific type of probability tree that enables you to plan about some kind of process. It is used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome.

- Supervised
- Classification/Regression
- Information Gain (IG)
- Entropy
- Gini Index

Tree in Data Structure

A decision tree is a particular type of probability tree that enables you to decide about some kind of process. It is used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome.



Explanations:

- **Root:** The top node (A) from which all other nodes descend.
- **Parent Node:** A node with child nodes (B is the parent of D).
- **Child Node:** A node that is a descendant of another node (D is a child of B).
- **Siblings:** Nodes with the same parent (F and G are siblings).
- **Leaf Node:** A node with no children (G, F, E, H, and I).
- **Sub-tree:** A part of the tree that can be considered its tree (the sub-tree rooted at D).

Problem Statement:

Class

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

Solving Logarithms:

1. **Direct Evaluation:**

If the argument of the log is a power of the base, directly evaluate.

- Example: $\log_2 8 = 3$ because $2^3 = 8$.

2. **Base Change Rule:**

To change the base of a logarithm, use the formula:

$\log_b a = \frac{\log_c a}{\log_c b}$, where c is any positive number.

- Example: $\log_2 10 = \frac{\log_{10} 10}{\log_{10} 2} = \frac{1}{\log_{10} 2}$.

3. **Product Rule:**

The log of a product is the sum of the logs:

$\log_b(MN) = \log_b M + \log_b N.$

- Example: $\log_2(3 \times 4) = \log_2 3 + \log_2 4$.

4. **Quotient Rule:**

The log of a quotient is the difference of the logs:

$$\log_b \left(\frac{M}{N}\right) = \log_b M - \log_b N.$$

- Example: $\log_2 \left(\frac{6}{3}\right) = \log_2 6 - \log_2 3.$

5. **Power Rule:**

The log of a power is the exponent times the log of the base:

$$\log_b(M^k) = k \cdot \log_b M.$$

- Example: $\log_2(8^2) = 2 \cdot \log_2 8 = 2 \cdot 3 = 6.$

6. **Inverse Rule:**

The base raised to the log of a number is just the number:

$$b^{\log_b a} = a.$$

- Example: $2^{\log_2 9} = 9.$

Relevant Equations

Wear Jacket?	
No	4 times
Yes	3 times

$$IG(Y,X) = E(Y) - E(Y|X)$$

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

E(Y) = Entropy Before Partition
E(Y|X) = Entropy After Partition
Target, E(Y) >> E(Y|X)

Entropy Before Partition:

Entropy of Wear Jacket:

= Entropy (4, 3)
= Entropy (– (Pi $\log_2$ Pi) + (– Pi $\log_2$ Pi))
= (-4/7 $\log_2$ 4/7) + (-3/7 $\log_2$ 3/7)
= (-.57 $\log_2$ .57) + (-.43 $\log_2$ .43)
= .985 (Entropy Before Partition)

Explanation:

- It starts with "Entropy (4, 3)," which indicates that there are two outcomes with 4 occurrences of the first outcome and 3 of the second.
- The formula for entropy used is $-\sum(p_i \log_2 p_i)$, where p_i represents the probabilities of the different outcomes.
- The probabilities are calculated as fractions of the total occurrences: $4/7$ and $3/7$.
- The entropy is then computed as $-((4/7)\log_2(4/7) + (3/7)\log_2(3/7))$.
- It simplifies further to $-0.57\log_2 .57$ for the first term and $-0.43\log_2 .43$ for the second term.
- The final numerical result is given as $.985$, which is labeled as "(Entropy Before Partition)."

Solving DT Classifier Algorithm

AI Quest STUDY MART

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Outlook
E (Outlook, Sunny) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812
E (Outlook, Cloudy) = -(2/3 $\log_2$ 2/3 + 1/3 $\log_2$ 1/3) = .918
Info Gain (S, Outlook) = E(S) – (4/7 * .812) – (3/7 * .918) = .985 – (4/7 * .812) – (3/7 * .918) = .127

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

Solving DT Classifier Algorithm

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Temperature
E (Temperature, Cold) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812
E (Temperature, Warm) = -(0/3 $\log_2$ 0/3 + 3/3 $\log_2$ 3/3) = 0.00
Info Gain (S, Temperature) = E(S) – (4/7 * .812) – (3/7 * 0) = .985 – (4/7 * .812) – (3/7*0) =.520

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

Solving DT Classifier Algorithm

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Routine
E (Routine, Indoor) = $-(1/4 \log_2 1/4 + 3/4 \log_2 3/4)$ = .812
E (Routine, Outdoor) = $-(2/3 \log_2 2/3 + 1/3 \log_2 1/3)$ = .918
Info Gain (S, Routine) = E(S) – (4/7*.812) – (3/7 * .918) = .985 – (4/7*.812) – (3/7 * .918) =.127

Problem Data Set

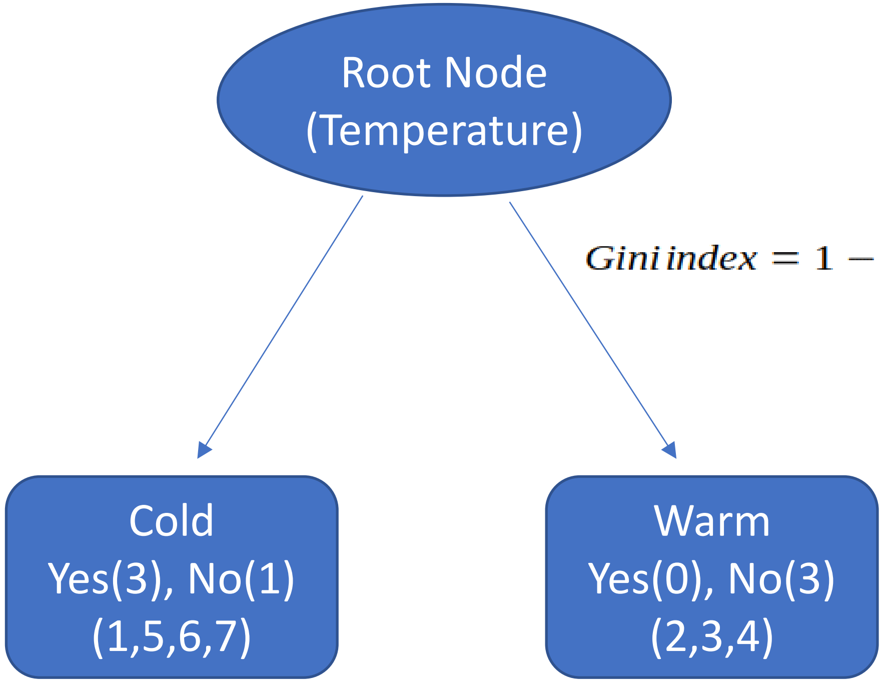
Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

Solving DT Classifier Algorithm

Root Node Selection Table

Outlook	Temperature	Routine
E (Outlook, Sunny) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812	E (Temperature, Cold) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812	E (Routine, Indoor) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812
E (Outlook, Cloudy) = -(2/3 $\log_2$ 2/3 + 1/3 $\log_2$ 1/3) = .918	E (Temperature, Warm) = -(0/3 $\log_2$ 0/3 + 3/3 $\log_2$ 3/3) = 0.00	E (Routine, Outdoor) = -(2/3 $\log_2$ 2/3 + 1/3 $\log_2$ 1/3) = .918
Info Gain (S, Outlook) = E(S) – (4/7 * .812) – (3/7 * .918) = .985 – (4/7 * .812) – (3/7 * .918) = .127	Info Gain (S, Temperature) = E(S) – (4/7 * .812) – (3/7 * 0) = .985 – (4/7 * .812) – (3/7*0) =.520	Info Gain (S, Routine) = E(S) – (4/7*.812) – (3/7 * .918) = .985 – (4/7*.812) – (3/7 * .918) =.127

Solving DT Classifier Algorithm

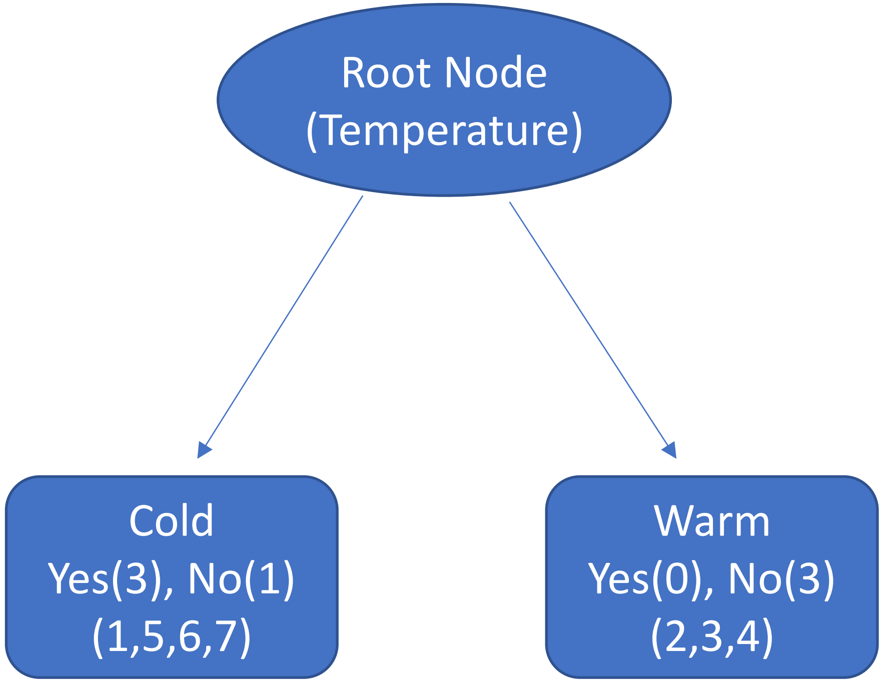


$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

Solving DT Classifier Algorithm



Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

Solving DT Classifier Algorithm

Entropy of New Subset:

S2 = Entropy (1,3)

$= \text{Entropy}\ (-(Pi \log_2 Pi) + (-Pi \log_2 Pi))$

$= (-1/4 \log_2 1/4) + (-3/4 \log_2 3/4)$

$= (-.25 \log_2 .25) + (-.75 \log_2 .75)$

= .812 (Entropy for New Subset)

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
~~2~~	~~Sunny~~	~~Warm~~	~~Outdoor~~	~~No~~
~~3~~	~~Cloudy~~	~~Warm~~	~~Indoor~~	~~No~~
~~4~~	~~Sunny~~	~~Warm~~	~~Indoor~~	~~No~~
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

Solving DT Classifier Algorithm

Entropy of New Subset:

S2 = Entropy (1,3)

= Entropy (– (Pi $\log_2$ Pi) + (– Pi $\log_2$ Pi))

= (-1/4 $\log_2$ 1/4) + (-3/4 $\log_2$ 3/4)

= (-.25 $\log_2$ .25) + (-.75 $\log_2$ .75)

= .812 (Entropy for New Subset)

Problem Data Set

Days	Outlook		Routine	Wear Jacket?
1	Sunny		Indoor	No
5	Cloudy		Indoor	Yes
6	Cloudy		Outdoor	Yes
7	Sunny		Outdoor	Yes

Solving DT Classifier Algorithm

E (Routine, Indoor) = $-(1/2 \log_2 ½ + 1/2 \log_2 1/2)$ $= 1$
E (Routine, Outdoor) = $-(2/2 \log_2 2/2 + 0/2 \log_2 0/2)$ $= 0$
Info Gain (S2, Routine) =   E(S2) – 2/4 * 1 – 2/4 *0   = .812 – 2/4 * 1 – 2/4 *0   = .312

Problem Data Set

Days	Outlook		Routine	Wear Jacket?
1	Sunny		Indoor	No
5	Cloudy		Indoor	Yes
6	Cloudy		Outdoor	Yes
7	Sunny		Outdoor	Yes

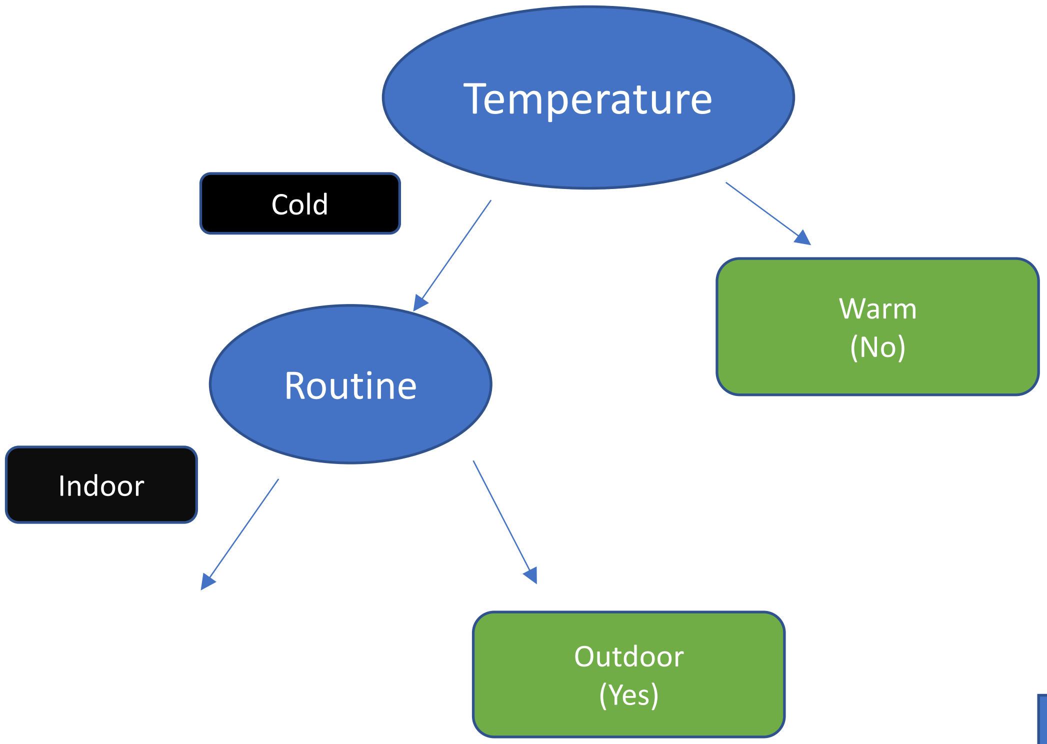
Solving DT Classifier Algorithm

E (Outlook, Sunny) = $-(1/2 \log_2 ½ + 1/2 \log_2 1/2)$ = 1
E (Outlook, Cloudy) = $-(2/2 \log_2 2/2 + 0/2 \log_2 0/2)$ = 0
Info Gain (S2, Outlook) =   E(S2) – 2/4 * 1 – 2/4 *0   = .812 – 2/4 * 1 – 2/4 *0   = .312

Problem Data Set

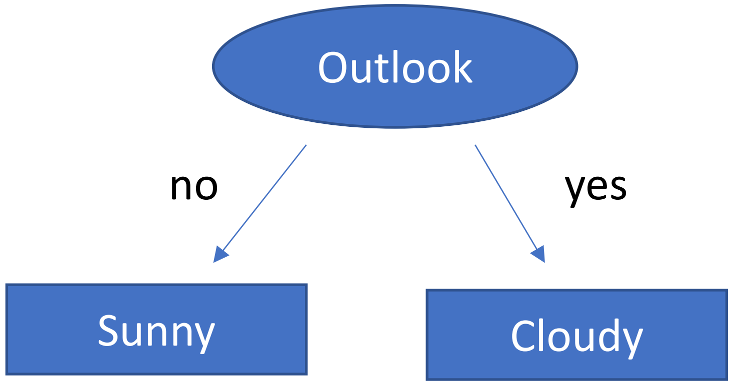
Days	Outlook		Routine	Wear Jacket?
1	Sunny		Indoor	No
5	Cloudy		Indoor	Yes
6	Cloudy		Outdoor	Yes
7	Sunny		Outdoor	Yes

Solving DT Classifier Algorithm



Sunny, Cold , Indoor= ??

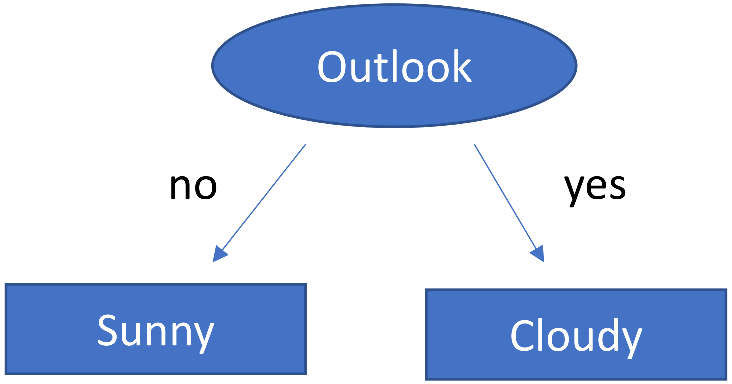
Solving DT Classifier Algorithm



Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
~~2~~	~~Sunny~~	~~Warm~~	~~Outdoor~~	~~No~~
~~3~~	~~Cloudy~~	~~Warm~~	~~Indoor~~	~~No~~
~~4~~	~~Sunny~~	~~Warm~~	~~Indoor~~	~~No~~
5	Cloudy	Cold	Indoor	Yes
~~6~~	~~Cloudy~~	~~Cold~~	~~Outdoor~~	~~Yes~~
~~7~~	~~Sunny~~	~~Cold~~	~~Outdoor~~	~~Yes~~

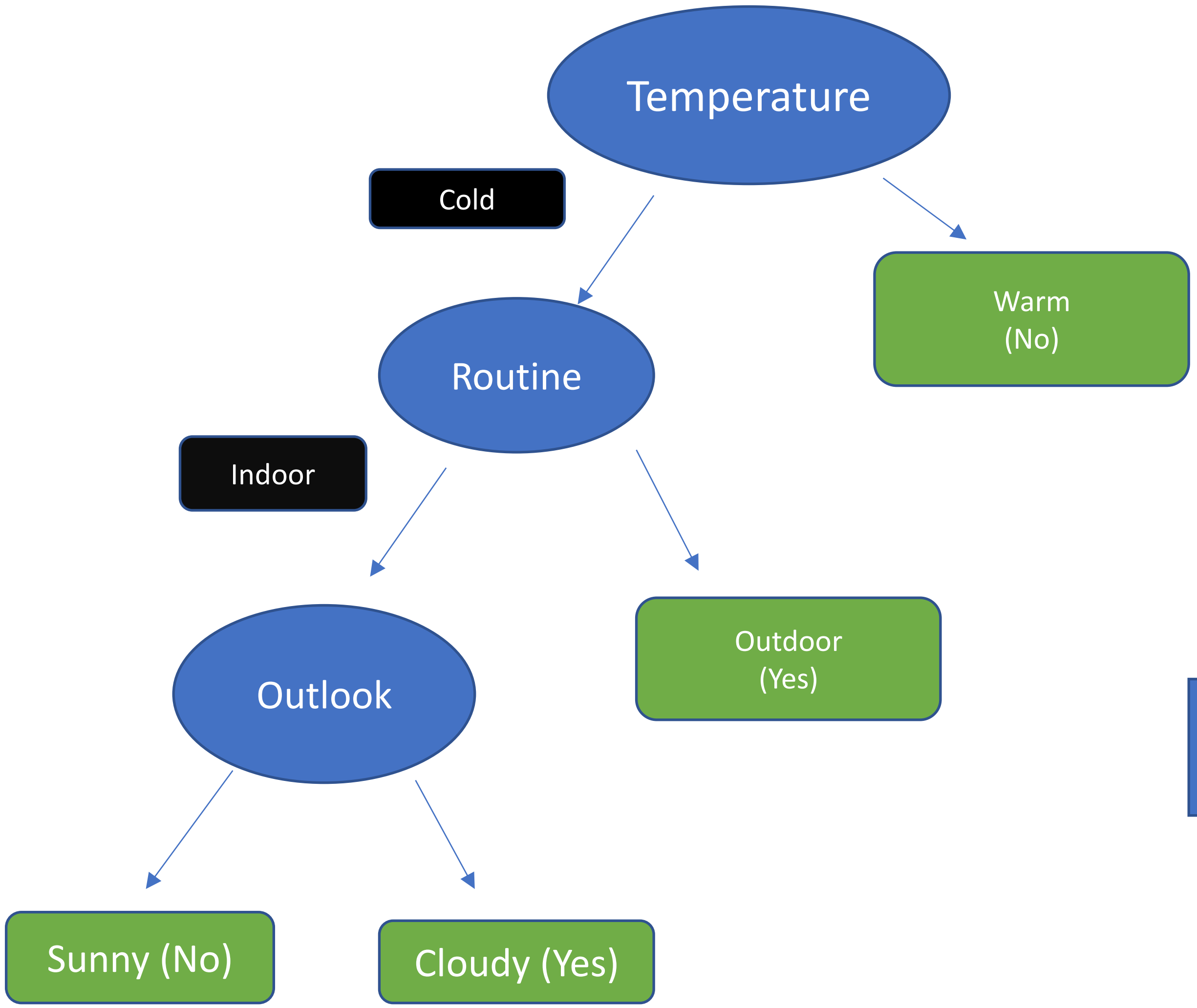
Solving DT Classifier Algorithm



Problem Data Set

Days	Outlook	Wear Jacket?
1	Sunny	No
5	Cloudy	Yes

Solving DT Classifier Algorithm



Sunny, Cold , Indoor= ??